

BEVFormer: 時空間変換器を用いたマルチカメラ画像からの鳥瞰表現の学習

Zhiqi Li^{1,2*}, Wenhui Wang^{2*}, Hongyang Li^{2*}, Enze Xie³, Chonghao Sima², Tong Lu¹, Yu Qiao², Jifeng Dai^{2✉}

¹Nanjing University ²Shanghai AI Laboratory ³The University of Hong Kong

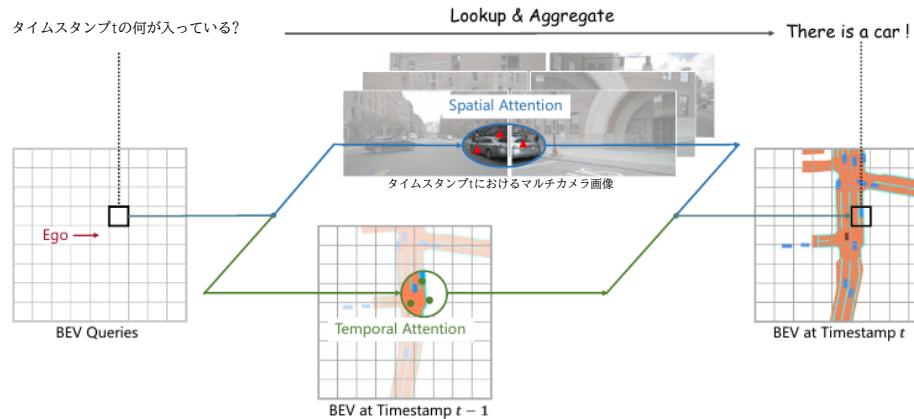


図1: マルチカメラ入力から鳥瞰(BEV)特徴を生成するために、TransformerとTemporal構造の両方を適用する自律走行のパラダイムであるBEVFormerを提案する。BEVFormerは、空間/時間空間を検索し、時空間情報を集約するためにクエリを活用する。

Abstract

自律走行システムにとって、3D検出やマルチカメラ画像に基づく地図分割を含む3D視覚認識タスクは不可欠である。本研究では、BEVFormerと呼ばれる新しいフレームワークを提示する。BEVFormerは、複数の自律走行知覚タスクをサポートするために、時空間変換器を用いて統一的なBEV表現を学習する。一言で言えば、BEVFormerは、あらかじめ定義されたグリッド状のBEVクエリを通じて、空間空間と時間空間と相互作用することで、空間情報と時間情報の両方を利用する。空間情報を集約するために、各BEVクエリがカメラビューをまたいで関心領域から空間特徴を抽出する空間交差注意を設計する。時間情報については、履歴BEV情報を再帰的に融合する時間的自己注意を提案する。我々のアプローチは、nuScenesテストセットにおいて、NDSメトリックの点で56.9%という新しい最先端を達成し、これは以前のベストアートよりも9.0ポイント高く、LiDARベースのベースラインの性能と同等である。さらに、BEVFormerが低視認性条件下での速度推定と物体の想起の精度を著しく向上させることを示す。コードは<https://github.com/zhiqi-li/BEVFormer>で公開されている。

*: Equal contribution. This work is done when Zhiqi Li is an intern at Shanghai AI Lab.
✉: Corresponding author.

1 Introduction

3D空間における知覚は、自律走行、ロボット工学など様々なアプリケーションにとって重要である。LiDARベースの手法[43, 20, 54, 50, 8]の目覚ましい進歩にもかかわらず、カメラベースのアプローチ[45, 32, 47, 30]は近年広く注目を集めている。展開のための低成本とは別に、カメラは、LiDARベースの対応するものと比較して、長距離距離のオブジェクトを検出し、ビジョンベースの道路要素(例えば、信号機、停留所)を識別するための望ましい利点を所有している。

自律走行における周囲のシーンの視覚認識は、複数のカメラから与えられる2Dキューから3Dバウンディングボックスやセマンティックマップを予測することが期待される。最も簡単な解決策は、単眼フレームワーク[45, 44, 31, 35, 3]とクロスカメラの後処理に基づくものである。このフレームワークの欠点は、異なるビューを別々に処理し、カメラ間で情報を捕捉できないため、性能と効率が低いことである[32, 47]。

単眼フレームワークの代替として、より統一的なフレームワークとして、マルチカメラ画像から全体的な表現を抽出する。鳥瞰図(BEV)は、物体の位置とスケールを明確に示すため、周囲のシーンを表現するのによく使われ、知覚や計画などの様々な自律走行タスクに適しています[29]。これまでのマップセグメンテーション手法はBEVの有効性を実証しているが[32, 18, 29]、BEVベースのアプローチは、3Dオブジェクト検出において他のパラダイムと比較して大きな利点を示していない[47, 31, 34]。その根本的な理由は、3Dオブジェクト検出タスクは、正確な3Dバウンディングボックス予測をサポートするために強力なBEV特徴を必要とするが、2D平面からBEVを生成することは非ポーズであるためである。BEV特徴を生成する一般的なBEVフレームワークは、深度情報[46, 32, 34]に基づいているが、このパラダイムは深度値の精度や深度分布に敏感である。このように、BEVベースの手法の検出性能は複合誤差[47]にさらされ、不正確なBEV特徴は最終的な性能を著しく損なう可能性がある。したがって、奥行き情報に依存せず、厳密に3次元事前分布に依存するのではなく、適忯的にBEV特徴を学習できるBEV生成手法を設計する動機付けとなる。Transformerは、注意メカニズムを使って価値ある特徴を動的に集約するもので、概念的に我々の要求に応えるものである。

知覚タスクを実行するためにBEV特徴を使用するもう一つの動機は、BEVが時間空間と空間空間を接続するための望ましい橋渡しであるということである。人間の視覚認識システムにとって、時間情報は物体の動きの状態を推測し、オクルージョンされた物体を識別する上で重要な役割を果たし、視覚分野の多くの研究がビデオデータを使用することの有効性を実証している[2, 27, 26, 33, 19]。しかし、既存の最先端のマルチカメラ3D検出手法は、時間情報を利用することはほとんどない。重要な課題は、自律走行は時間的に重要であり、シーン内のオブジェクトは急速に変化することである。したがって、クロスタイムスタンプのBEV特徴を単純に積み重ねることは、余分な計算コストと干渉情報をもたらし、理想的でないかもしれない。リカレントニューラルネットワーク(RNN)[17, 10]に触発され、我々はBEV特徴を利用して、過去から現在までの時間情報をリカレントに配信する。

この目的のために、我々はBEVFormerと呼ばれる変換器ベースの鳥瞰(BEV)エンコーダを提示し、マルチビューカメラからの時空間特徴と履歴BEV特徴を効果的に集約することができる。BEVFormerから生成されたBEV特徴量は、3D物体検出や地図分割など、複数の3D知覚タスクを同時にサポートすることができ、自律走行システムにとって有用である。図1に示すように、我々のBEVFormerは3つの主要な設計を含んでいる。(1)注意メカニズムを介して空間的特徴と時間的特徴を柔軟に融合するグリッド状のBEVクリエイタ、(2)マルチカメラ画像から空間的特徴を集約する空間的交差注意モジュール、(3)履歴BEV特徴から時間的情報を抽出する時間的自己注意モジュールであり、これにより移動物体の速度推定と重く隠蔽された物体の検出が有益になる。verhead。BEVFormerによって生成された統一的な特徴により、このモデルは、エンドツーエンドの3Dオブジェクト検出とマップセグメンテーションのために、Deformable DETR [56]やマスクデコーダ[22]などの異なるタスク固有のヘッドと協調することができる。

我々の主な貢献は以下の通りである：

- BEVFormerは、マルチカメラおよび/またはタイムスタンプの入力をBEV表現に投影する時空間変換エンコーダである。統一されたBEV特徴により、我々のモデルは、3D検出や地図分割を含む複数の自律走行知覚タスクを同時にサポートすることができる。

- 学習可能なBEVクエリを、空間的交差注意層と時間的自己注意層とともに設計し、それぞれ交差カメラからの空間的特徴と履歴BEVからの時間的特徴を検索し、それらを統一的なBEV特徴に集約する。
- 提案するBEVFormerを、nuScenes [4]やWaymo [40]などの複数の困難なベンチマークで評価する。我々のBEVFormerは、先行技術と比較して一貫して性能向上を達成している。例えば、同等のパラメータと計算オーバヘッドの下で、BEVFormerはnuScenesテストセットで56.9%のNDSを達成し、以前の最良検出手法DETR3D [47]を9.0ポイント上回った(56.9%対47.9%)。地図セグメンテーションタスクにおいても、最も困難な車線セグメンテーションにおいて、Lift-Splat [32]よりも5.0ポイント以上高い、最先端の性能を達成した。我々は、この単純で強力なフレームワークが、以下の3D知覚タスクのための新しいベースラインとして役立つことを期待している。

2 関連研究

2.1 トランスフォーマーに基づく2次元知覚

最近、検出とセグメンテーションのタスクを再定式化するために、変換器を使用することが新しい傾向となっている[7, 56, 22]。

DETR [7]は、オブジェクトクエリのセットを使用して、クロスアテンションデコーダによる検出結果を直接生成する。しかし、DETRの主な欠点は学習時間が長いことである。Deformable DETR [56]は、変形可能な注意を提案することで、この問題を解決する。DETRのバニラグローバルアテンションとは異なり、変形可能なアテンションは局所的な関心領域と相互作用し、各参照点近傍のK点のみをサンプリングしてアテンション結果を計算するため、高い効率性と学習時間の大規模な短縮を実現する。変形可能な注意メカニズムは次式で計算される：

$$\text{DeformAttn}(q, p, x) = \sum_{i=1}^{N_{\text{head}}} \mathcal{W}_i \sum_{j=1}^{N_{\text{key}}} \mathcal{A}_{ij} \cdot \mathcal{W}'_i x(p + \Delta p_{ij}), \quad (1)$$

ここで、 q , p , x はそれぞれクエリ、参照点、入力特徴量を表す。 i は注意ヘッドを、 N_{head} は注意ヘッドの総数を表す。 j はサンプリングされたキーをインデックスし、 N_{key} は各ヘッドのサンプリングされたキー番号の合計である。 $\mathcal{W}_i \in \{\mathbb{R}\}^{\{C \times (C/H \times H)\}}$ と $\mathcal{W}'_i \in \{\mathbb{R}\}^{\{C \times (H \times W)\}}$ は学習可能な重みであり、 C は特徴次元である。 $\mathcal{A}_{ij} \in \{0, 1\}$ は予測注目重みであり、 $\sum_{j=1}^{N_{\text{key}}} \mathcal{A}_{ij} = 1$ で正規化される。 $\Delta p_{ij} \in \mathbb{R}^2$ は参照点 p に対する予測オフセットである。 $x(p + \Delta p_{ij})$ は位置 $p + \Delta p_{ij}$ の特徴量を表し、Dai ら[12]と同様にバイリニア補間により抽出される。本研究では、変形可能な注意を3次元知覚タスクに拡張し、空間情報と時間情報の両方を効率的に集約する。

2.2 カメラベースの3D知覚

これまでの3D知覚手法では、3D物体検出や地図分割のタスクを独立して行うのが一般的であった。3D物体検出タスクでは、初期の手法は2D検出手法[1, 28, 49, 39, 53]に似ており、通常2Dバウンディングボックスに基づいて3Dバウンディングボックスを予測する。Wangら[45]は、高度な2D検出器FCOS[41]に従い、各オブジェクトの3Dバウンディングボックスを直接予測する。DETR3D[47]は、2D画像に学習可能な3Dクエリを投影し、NMSの後処理を行わずに、エンドツーエンドの3Dバウンディングボックス予測のために対応する特徴をサンプリングする。もう一つの解決策は、画像特徴をBEV特徴に変換し、トップダウンビューから3Dバウンディングボックスを予測することである。深度推定[46]やカテゴリカルな深度分布[34]から深度情報を用いて、画像特徴をBEV特徴に変換する手法がある。OFT [36]とImVoxelNet [37]は、あらかじめ定義されたボクセルを画像特徴に投影し、シーンのボクセル表現を生成する。最近、M² BEV [48]は、BEV特徴に基づく複数の知覚タスクを同時に実行することの実現可能性をさらに探求した。

実際、マルチカメラ特徴からBEV特徴を生成することは、マップセグメンテーションタスクにおいてより広範に研究されている[32, 30]。透視図を逆透視図法(IPM)[35, 5]によってBEVに変換する簡単な方法がある。また、Lift-Splat[32]は、奥行き分布に基づいてBEV特徴量を生成する。方法[30, 16, 9]は、多層ペーセプトロンを利用して、透視図からBEVへの変換を学習する。PYVA [51]は、正面から見た単眼画像をBEVに変換するクロスピュー変換器を提案しているが、このパラダイムは

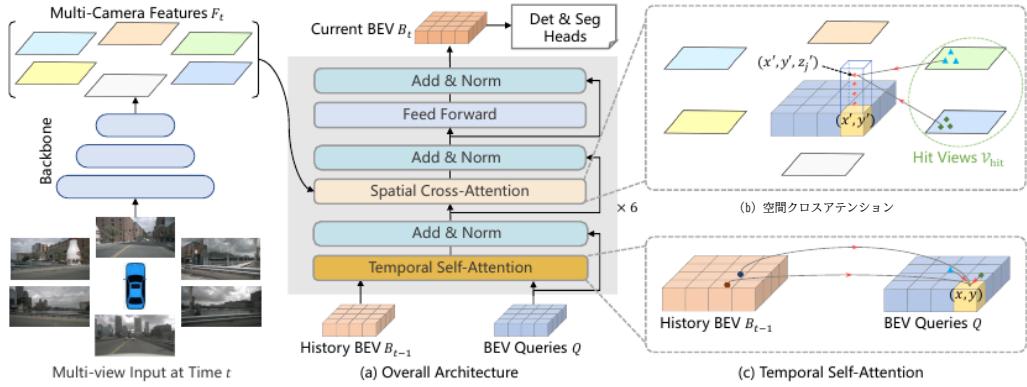


図2:BEVFormerの全体アーキテクチャ。(a) BEVFormerのエンコーダ層は、グリッド状のBEVクエリ、時間的自己注意、空間的交差注意を含む。(b) 空間交差注意では、各BEVクエリは関心領域の画像特徴のみと相互作用する。(c)時間的自己アテンションでは、各BEVクエリは2つの特徴と相互作用する:現在のタイムスタンプのBEVクエリと前のタイムスタンプのBEV特徴。

グローバルアテンションメカニズムの計算コストのため、マルチカメラ特徴を融合する[42]。空間情報を加えて、先行研究[18, 38, 6]では、複数のタイムスタンプからBEV特徴を積み重ねることで、時間情報も考慮している。BEV特徴量の積み重ねは、利用可能な時間情報を一定時間内に制約し、余分な計算コストをもたらす。本研究では、提案する時空間変換器を用いて、空間的手がかりと時間的手がかりの両方を考慮して現在時刻のBEV特徴を生成し、RNN法により過去のBEV特徴から時間情報を得るが、計算量はほとんどかからない。

3 BEVFormer

マルチカメラ画像特徴を鳥瞰(BEV)特徴に変換することで、様々な自律走行知覚タスクのための統一された周辺環境表現を提供することができる。本研究では、マルチビューカメラからの時空間特徴と、注意メカニズムを介した履歴BEV特徴を効率的に集約することができる、BEV生成のための新しい変換器ベースのフレームワークを提示する。

3.1 全体アーキテクチャ

図2に示すように、BEVFormerは6つのエンコーダ層を持ち、各エンコーダ層は、BEVクエリ、空間交差注意、時間自己注意という3つのテーラードデザインを除いて、従来の変換器[42]の構造に従っている。具体的には、BEVクエリはグリッド状の学習可能なパラメータであり、注意メカニズムを介してマルチカムビューからBEV空間の特徴をクエリするよう設計されている。空間的交差注意と時間的自己注意は、BEVクエリで動作する注意層であり、BEVクエリに従って、履歴BEVからの時間的特徴だけでなく、マルチカム画像からの空間的特徴を検索し、集約するために使用される。

推論中、タイムスタンプ t で、マルチカム画像をバックボーンネットワーク(例えば、ResNet-101[15])に供給し、異なるカムラビュの特徴量 $F_t = \{F_t^i\}_{i=1}^{N_{view}}$ を得る、ここで F_t^i は i 番目のビュの特徴量、 N_{view} はカムラビュの総数である。同時に、BEV特徴量 B_{t-1} を事前タイムスタンプ $t-1$ に保存した。各エンコーダ層では、まずBEVクエリ Q を用いて、事前BEV特徴量 B_{t-1} から時間的自己アテンションを介して時間情報をクエリする。次に、BEVクエリ Q を用いて、空間クロスアテンションを介して、マルチカム特徴 F_t から空間情報を問い合わせる。フィードフォワードネットワーク[42]の後、エンコーダ層は洗練されたBEV特徴を出力し、これは次のエンコーダ層の入力となる。6層のエンコーダ層を重ねた後、現在のタイムスタンプ t における統一BEV特徴量 B_t を生成する。BEV特徴量 B_t を入力として、3D検出ヘッドとマップ分割ヘッドは、3Dバウンディングボックスやセマンティックマップなどの知覚結果を予測する。

3.2 BEVクエリ

グリッド形状の学習可能なパラメータ群 $Q \in R^{H \times W \times C}$ をBEVFormerのクエリとして定義する。具体的には、 Q の $p=(x, y)$ に位置するクエリ $Q_p \in R^{1 \times C}$ が、BEV平面上の対応するグリッドセル領域を担当する。BEV平面上の各グリッドセルは、実世界の s メートルのサイズに対応する。BEV特徴量の中心は、デフォルトではエゴ・カーの位置に対応する。一般的な手法[14]に従い、BEVFormerに入力する前に、BEVクエリ Q に学習可能な位置埋め込みを追加する。

3.3 空間的交差注意

マルチカメラ3D知覚の入力スケールが大きい(N_{view} 個のカメラビューを含む)ため、バニラマルチヘッドアテンション[42]の計算コストが非常に高い。そこで、各BEVクエリ Q_p がカメラビューをまたいでその関心領域とのみ相互作用する、リソース効率の良い注意層である変形可能注意[56]に基づく空間交差注意を開発する。しかし、変形可能な注意はもともと2D知覚のために設計されているため、3Dシーンにはいくつかの調整が必要である。

図2(b)に示すように、まずBEV平面上の各クエリを柱状のクエリ[20]に持ち上げ、柱から N_{ref} 個の3D参照点をサンプリングし、これらの点を2Dビューに投影する。1つのBEVクエリに対して、投影された2Dポイントは一部のビューにしか該当せず、他のビューはヒットしない。ここで、ヒットビューを V_{hit} と呼ぶ。その後、これらの2次元点をクエリ Q_p の参照点とみなし、これらの参照点の周りのヒットビュー V_{hit} から特徴をサンプリングする。最後に、空間的交差注意の出力として、サンプリングされた特徴量の加重和を実行する。空間交差注意(SCA)のプロセスは、以下のように定式化できる：

$$SCA(Q_p, F_t) = \frac{1}{|V_{hit}|} \sum_{i \in V_{hit}} \sum_{j=1}^{N_{ref}} \text{DeformAttn}(Q_p, \mathcal{P}(p, i, j), F_t^i), \quad (2)$$

ここで、 i はカメラビュー、 j は参照点、 N_{ref} は各BEVクエリの総参照点である。 F_t^i は i 番目のカメラビューの特徴量である。各BEVクエリ Q_p に対して、プロジェクト関数 $P(p, i, j)$ を用いて、 i 番目のビュー画像上の j 番目の参照点を得る。

次に、投影関数 P からビュー画像上の参照点を得る方法を紹介する。まず、 Q の $p=(x, y)$ に位置するクエリ Q_p に対応する実世界の位置 (x^0, y^0) を式3のように計算する。

$$x' = (x - \frac{W}{2}) \times s; \quad y' = (y - \frac{H}{2}) \times s, \quad (3)$$

ここで、 H, W はBEVクエリの空間形状、 s はBEVのグリッドの解像度の大きさ、 (x^0, y^0) はエゴカーの位置が原点となる座標である。3次元空間では、 (x^0, y^0) に位置する物体は z 軸上の z^0 の高さに現れる。そこで、アンカーの高さ $\{z_j^0\}_{j=1}^{N_{ref}}$ のセットをあらかじめ定義し、異なる高さに現れた手がかりを捉えることができるようになる。このようにして、各クエリ Q_p に対して、3次元参照点 $(x^0, y^0, z_j^0)_{j=1}^{N_{ref}}$ の柱を得る。最後に、カメラの投影行列を通して、3D参照点を異なる画像ビューに投影します：

$$\begin{aligned} \mathcal{P}(p, i, j) &= (x_{ij}, y_{ij}) \\ \text{where } z_{ij} \cdot [x_{ij} \ y_{ij} \ 1]^T &= T_i \cdot [x' \ y' \ z'_j \ 1]^T. \end{aligned} \quad (4)$$

ここで、 $P(p, i, j)$ は j 番目の3次元点 (x^0, y^0, z_j^0) から投影された i 番目のビュー上の2次元点、 $T_i \in R^{3 \times 4}$ は i 番目のカメラの既知の投影行列である。

3.4 時間的自己アテンション

空間情報に加えて、時間情報も視覚系が周囲の環境を理解するために重要である[27]。例えば、時間的な手がかりがない静止画像から、動いている物体の速度を推測したり、高度に隠蔽された物体を検出したりすることは困難である。この問題に対処するため、履歴BEV特徴を取り込むことで現在の環境を表現できる時間的自己注意を設計する。

現在のタイムスタンプ t におけるBEVクエリ Q と、タイムスタンプ $t-1$ に保存された履歴BEV特徴 B_{t-1} が与えられたとき、まず B_{t-1} をエゴモーションに従って Q に整列させ、同じグリッドの特徴が同じ実世界の位置に対応するようにする。ここで、整列履歴BEV特徴量 B_{t-1} を B_{t-1}^0 とする。しかし、時刻 $t-1$ から t まで、可動物体は様々なオフセットで実世界を移動している。異なる時間のBEV特徴間の同じオブジェクトの正確な関連付けを構築することは困難である。そこで、この特徴間の時間的なつながりを時間的自己注意(TSA)層を通してモデル化する：

$$\text{TSA}(Q_p, \{Q, B'_{t-1}\}) = \sum_{V \in \{Q, B'_{t-1}\}} \text{DeformAttn}(Q_p, p, V), \quad (5)$$

ここで、 Q_p は $p = (x, y)$ に位置するBEVクエリを表す。また、バニラ変形可能注意とは異なり、時間的自己注意におけるオフセット Δp は、 Q と B_{t-1}^0 の連結によって予測される。特に、各シーケンスの最初のサンプルについて、時間的自己注意は時間情報のない自己注意に縮退し、BEV特徴 $\{Q, B_{t-1}^0\}$ を重複BEVクエリ $\{Q, Q\}$ に置き換える。

$[18, 38, 6]$ のBEVを単純に積み重ねるよりも、我々の時間的自己注意は長い時間的依存性をより効果的にモデル化できる。BEVFormerは、複数のBEV特徴を積み重ねるのではなく、前のBEV特徴から時間情報を抽出するため、計算コストが少なくて済み、邪魔な情報も少なくて済む。

3.5 BEV機能の応用

BEV特徴量 $B_t \in R^{H \times W \times C}$ は、様々な自律走行知覚タスクに利用できる汎用性の高い2次元特徴マップであるため、2次元知覚手法[56, 22]に基づいて、3次元物体検出とマップ分割タスクヘッドを若干修正しながら開発することができる。

3D物体検出のために、2D検出器Deformable DETR [56]に基づいて、エンドツーエンドの3D検出ヘッドを設計する。修正点としては、デコーダの入力としてシングルスケールのBEV特徴量 B_t を用いること、2次元のバウンディングボックスではなく3次元のバウンディングボックスと速度を予測すること、3次元のバウンディングボックス回帰を監視するために L_1 損失のみを用いることなどがある。検出ヘッドを用いることで、我々のモデルはNMSの後処理なしに、3Dバウンディングボックスと速度をエンドツーエンドで予測することができる。

地図セグメンテーションのために、2Dセグメンテーション手法Panoptic SegFormer [22]に基づいて、地図セグメンテーションヘッドを設計する。BEVに基づくマップセグメンテーションは、基本的に一般的なセマンティックセグメンテーションと同じであるため、[22]のマスクデコーダとクラス固定クエリを利用し、車、車両、道路(走行可能領域)、車線などの各セマンティックカテゴリを対象とする。

3.6 実装の詳細

学習段階。タイムスタンプ t の各サンプルについて、過去2秒間の連続したシーケンスからさらに3つのサンプルをランダムにサンプリングし、このランダムサンプリング戦略によって自我運動の多様性を増強することができる[57]。これら4つのサンプルのタイムスタンプを $t-3, t-2, t-1, t$ とする。最初の3つのタイムスタンプのサンプルに対して、BEV特徴 $\{B_{t-3}, B_{t-2}, B_{t-1}\}$ を再帰的に生成する役割を担っており、このフェーズでは勾配は必要ない。タイムスタンプ $t-3$ の最初のサンプルでは、BEVの特徴は過去に存在せず、時間的自己注意は自己注意に縮退する。時刻 t において、モデルはマルチカメラ入力と事前BEV特徴 B_{t-1} の両方に基づいてBEV特徴 B_t を生成し、 B_t が4つのサンプルを横切る時間的・空間的手がかりを含むようにする。最後に、BEV特徴量 B_t を検出ヘッドとセグメンテーションヘッドに入力し、対応する損失関数を計算する。

推論フェーズ。推論フェーズでは、ビデオシーケンスの各フレームを時系列に評価する。前のタイムスタンプのBEV特徴量を保存し、次のタイムスタンプに使用する。このオンライン推論戦略は時間効率が良く、実用的なアプリケーションと一致する。我々は時間情報を利用しているが、推論速度は他の手法[45, 47]と同等である。

4 Experiments

4.1 Datasets

我々は、nuScenesデータセット[4]とWaymoオープンデータセット[40]という2つの困難な公共自律走行データセットで実験を行う。

nuScenesデータセット[4]は、それぞれおよそ20秒の1000シーンを含み、キーサンプルは2Hzでアノテーションされている。各サンプルは6台のカメラによるRGB画像で構成され、360°の水平FOVを持つ。検出タスクでは、10カテゴリから1.4Mの注釈付き3Dバウンディングボックスがある。BEVセグメンテーションタスクは[32]の設定に従う。このデータセットはまた、検出タスクの公式な評価指標を提供する。nuScenesの平均平均精度(mAP)は、予測結果とグランドトゥルースを一致させるために、3D Intersection over Union(IoU)ではなく、地上面上の中心距離を用いて計算される。nuScenesメトリクスには、翻訳、スケール、方向、速度、属性エラーをそれぞれ測定するためのATE、ASE、AOE、AVE、AAEを含む5種類の真陽性メトリクス(TPメトリクス)も含まれている。nuScenesはまた、nuScenes検出タスクの全ての側面を捉るために、 $NDS = \frac{1}{10} [5mAP + \sum_{mTP \in TP} (1 - \min(1, mTP))]$ としてnuScenes検出スコア(NDS)を定義する。

Waymo Open Dataset [40]は、798の学習シーケンスと202の検証シーケンスを持つ大規模な自律走行データセットである。Waymoが提供する各フレームの5つの画像は、水平FOVが約252°しかないが、提供された注釈付きラベルはエゴ・カーの周囲360°であることに注意。訓練セットと検証セットとの画像にも見えないバウンディングボックスを削除する。Waymo Open Datasetは大規模かつ高速であるため[34]、学習シーケンスから5thフレーム毎にサンプリングすることで、学習分割のサブセットを使用し、車両カテゴリのみを検出する。WaymoデータセットのmAPを計算するために、3D IoUの閾値0.5と0.7を使用する。

4.2 実験設定

先行手法[45, 47, 31]に従い、2種類のバックボーンを採用する：FCOS3D[45]チェックポイントから初期化したResNet101-DCN[15, 12]と、DD3D[31]チェックポイントから初期化したVoVnet-99[21]である。デフォルトでは、FPN [23]の出力マルチスケール特徴量を利用し、サイズは $1 / 16, 1 / 32, 1 / 64$ 、次元は $C = 256$ とする。nuScenesでの実験では、BEVクエリのデフォルトサイズは 200×200 、知覚範囲はX軸とY軸で[-5 1.2m, 51.2m]、BEVのグリッドの解像度sのサイズは0.512mである。BEVクエリには学習可能な位置埋め込みを採用する。BEVエンコーダは6つのエンコーダ層を含み、各層で常にBEVクエリを洗練する。各エンコーダ層の入力BEV特徴量 B_{t-1} は同じであり、勾配を必要としない。各ローカルクエリに対して、変形可能注意メカニズムによって実装された空間交差注意モジュールの間、それは3D空間における異なる高さを持つ $N_{ref} = 4$ 個のターゲット点に対応し、事前に定義された高さアンカーは-5メートルから3メートルまで一様にサンプリングされる。2Dビュー特徴上の各基準点について、各ヘッドについてこの基準点の周囲に4つのサンプリング点を使用する。デフォルトでは、24エポック、学習率 2×10^{-4} でモデルを学習する。

Waymoでの実験では、いくつかの設定を変更する。Waymoのカメラシステムはエゴ・カー[40]周辺のシーン全体をキャプチャできないため、BEVクエリのデフォルトの空間形状は 300×220 であり、知覚範囲はX軸が[-35.0m, 75.0m]、Y軸が[-75.0m, 75.0m]である。各軸の解像度 s のサイズは 0.5m である。自車両はBEVの(70, 150)にある。

ベースライン。タスクヘッドの影響を排除し、他のBEV生成方法を公平に比較するために、VPN [30]とLift-Spl at [32]を使用して、我々のBEVFormerを置き換え、タスクヘッドと他の設定を同じに保つ。また、BEVFormer-Sと呼ばれる静的モデルに適応させるために、履歴BEV特徴を用いずに、時間的自己注意をバニラ的自己注意に調整する。

4.3 3次元物体検出結果

我々は、従来の最先端の3D物体検出手法と公平に比較するために、検出ヘッドのみで検出タスクで我々のモデルを訓練する。Tab. 1とTab. 2では、nuScenesのテストとvalの分割に関する主な結果を報告する。我々の手法は、公正な学習戦略と同等のモデルスケールの下で、valセットで9.2ポイント(51.7% NDS対42.5% NDS)以上、以前の最良手法DETR3D [47]を上回る。テストセットにおいて、我々のモデルはベルとホイッスルなしで56.9%のNDSを達成し、DETR3D(47.9%のNDS)より9.0ポイント高い。

表1: nuScenesテストセットでの3D検出結果。* VoVNet-99 (V2-99) [21]は、追加データ[31]を用いて深度推定タスクで事前学習されたことに注意。”BEVFormer-S”はBEVエンコーダの時間情報を活用しない。”L”はLiDAR、“C”はCameraを示す。

Method	Modality		Backbone		NDS↑	mAP↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
SSN [55]	L	-	0.569	0.463	-	-	-	-	-	-	-
CenterPoint-Voxel [52]	L	-	0.655	0.580	-	-	-	-	-	-	-
PointPainting [43]	L&C	-	0.581	0.464	0.388	0.271	0.496	0.247	0.111		
FCOS3D [45]	C	R101	0.428	0.358	0.690	0.249	0.452	1.434	0.124		
PGD [44]	C	R101	0.448	0.386	0.626	0.245	0.451	1.509	0.127		
BEVFormer-S	C	R101	0.462	0.409	0.650	0.261	0.439	0.925	0.147		
BEVFormer	C	R101	0.535	0.445	0.631	0.257	0.405	0.435	0.143		
DD3D [31]	C	V2-99*	0.477	0.418	0.572	0.249	0.368	1.014	0.124		
DETR3D [47]	C	V2-99*	0.479	0.412	0.641	0.255	0.394	0.845	0.133		
BEVFormer-S	C	V2-99*	0.495	0.435	0.589	0.254	0.402	0.842	0.131		
BEVFormer	C	V2-99*	0.569	0.481	0.582	0.256	0.375	0.378	0.126		

表2: nuScenes val setでの3D検出結果。”C”はカメラを示す。

Method	Modality	Backbone	NDS↑	mAP↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
FCOS3D [45]	C	R101	0.415	0.343	0.725	0.263	0.422	1.292	0.153
PGD [44]	C	R101	0.428	0.369	0.683	0.260	0.439	1.268	0.185
DETR3D [47]	C	R101	0.425	0.346	0.773	0.268	0.383	0.842	0.216
BEVFormer-S	C	R101	0.448	0.375	0.725	0.272	0.391	0.802	0.200
BEVFormer	C	R101	0.517	0.416	0.673	0.274	0.372	0.394	0.198

我々の手法は、SSN(56.9% NDS)[55]やPointPainting(58.1% NDS)[43]のようないくつかのLiDARベースのベースラインに匹敵する性能さえ達成できる。

これまでのカメラベースの手法[47, 31, 45]では、速度をほとんど推定できず、本手法は、時間情報がマルチカメラ検出の速度推定に重要な役割を果たすことを実証している。BEVFormerの平均速度誤差(mAVE)は、テストセットで0.378m/sであり、他のカメラベースの手法を大差で上回り、LiDARベースの手法の性能に近づいている[43]。

また、Tab. 3に示すように、Waymoでの実験も行っている。3. 34]に従い、IoU基準を0.7と0.5として車両カテゴリを評価する。また、カメラベースの手法ではIoUベースのメトリクスが難しすぎるため、nuScenesメトリクスを採用して結果を評価する。Waymoで結果を報告したカメラベースの研究がいくつかあるため、比較のためにDETR3Dの公式コードもWaymoで実験を行うために使用する。BEVFormerは、IoU基準0.5でのLEVEL_1とLEVEL_2の難易度において、ヘディング情報付き平均精度(APH)[40]でDETR3Dを6.0%と2.5%上回ることが観察できる。nuScenesメトリクスでは、BEVFormerは3.2%のNDSと5.2%のAPのマージンでDETR3Dを上回る。また、BEVFormerとCaDNN[34](Waymoデータセットでの結果を報告した単眼3D検出手法)を比較するために、フロントカメラでの実験も行った。BEVFormerは、IoU基準0.5のLEVEL_1とLEVEL_2の難易度において、APH13.3%と11.2%でCaDNNを上回った。

4.4 マルチタスク知覚結果

複数のタスクに対するモデルの学習能力を検証するために、検出ヘッドとセグメンテーションヘッドの両方を用いてモデルを学習させ、その結果をTab. 4. 同じ設定で異なるBEVエンコーダを比較した場合、BEVFormerは道路セグメンテーションの結果がBEVFormer-Sと同等である以外は、全てのタスクでより高い性能を達成している。例えば、共同学習では、BEVFormerはLift-Splat* [32]を、デテーションタスクで11.0ポイント(NDS52.0%対NDS41.0%)、車線分割で5.6ポイント(23.9%対18.3%)上回った。マルチタスク学習は、個別に学習するタスクと比較して、バックボーンやBEVエンコーダを含むより多くのモジュールを共有することで、計算コストを節約し、推論時間を短縮する。

表3: Waymo評価指標とnuScenes評価指標におけるWaymo valセットの3D検出結果。「L1」と「L2」は、Waymo [40]の「LEVEL_1」と「LEVEL_2」の難しさを指す。*: 前面カメラのみを使用し、前面カメラの視野(50.4°)内のオブジェクトラベルのみを考慮する。†: ATEとAAを1としてNDSスコアを算出する。「L」はLiDAR、「C」はCameraをそれぞれ示す。

Method	Modality	Waymo Metrics				Nuscenes Metrics			
		$\text{IoU} = 0.5$	$\text{IoU} = 0.7$	$\text{IoU} = 0.5$	$\text{IoU} = 0.7$	NDS \uparrow	AP \uparrow	ATE \downarrow	ASE \downarrow
PointPillars [20]	L	0.866	0.801	0.638	0.557	0.685	0.838	0.143	0.132
DETR3D [47]	C	0.220	0.216	0.055	0.051	0.394	0.388	0.741	0.156
BEVFormer	C	0.280	0.241	0.061	0.052	0.426	0.440	0.679	0.157
CaDNN* [34]	C	0.175	0.165	0.050	0.045	-	-	-	-
BEVFormer*	C	0.308	0.277	0.077	0.069	-	-	-	-

表4: nuScenes val setにおける3D検出とマップセグメンテーションの結果。セグメンテーションと検出タスクの学習を合同で行う場合と行わない場合の比較。*: 比較のために、VPN [30]とLift-Splat [32]を使用して、我々のBEVエンコーダを置き換える、タスクヘッドは同じである。†: 彼らの論文の結果

Method	Task Head		3D Detection		BEV Segmentation (IoU)			
	Det	Seg	NDS \uparrow	mAP \uparrow	Car	Vehicles	Road	Lane
Lift-Splat [†] [32]	✗	✓	-	-	32.1	32.1	72.9	20.0
FIERY [†] [18]	✗	✓	-	-	-	38.2	-	-
VPN* [30]	✓	✗	0.333	0.253	-	-	-	-
VPN*	✗	✓	-	-	31.0	31.8	76.9	19.4
VPN*	✓	✓	0.334	0.257	36.6	37.3	76.0	18.0
Lift-Splat*	✓	✗	0.397	0.348	-	-	-	-
Lift-Splat*	✗	✓	-	-	42.1	41.7	77.7	20.0
Lift-Splat*	✓	✓	0.410	0.344	43.0	42.8	73.9	18.3
BEVFormer-S	✓	✗	0.448	0.375	-	-	-	-
BEVFormer-S	✗	✓	-	-	43.1	43.2	80.7	21.3
BEVFormer-S	✓	✓	0.453	0.380	44.3	44.4	77.6	19.8
BEVFormer	✓	✗	0.517	0.416	-	-	-	-
BEVFormer	✗	✓	-	-	44.8	44.8	80.1	25.7
BEVFormer	✓	✓	0.520	0.412	46.8	46.7	77.5	23.9

本論文では、我々のBEVエンコーダによって生成されたBEV特徴が、異なるタスクにうまく適応できること、そして、マルチタスクヘッドを用いたモデル学習が、検出タスクと車両セグメンテーションにおいてさらに優れた性能を発揮することを示す。しかし、マルチタスク学習においてネガティブransformer [11, 13]と呼ばれる一般的な現象である道路や車線のセグメンテーションでは、共同学習したモデルは個別に学習したモデルほど性能が高くない。

4.5 アブレーション研究

異なるモジュールの効果を調べるために、検出ヘッドを持つnuScenes val setでアブレーション実験を行う。より多くのアブレーション研究は付録にある。

空間的クロスアテンションの有効性空間的交差注意の効果を検証するために、BEVFormer-Sを用いて、時間情報の干渉を排除するアブレーション実験を行い、その結果をTab. 5. デフォルトの空間的交差注意は変形可能な注意に基づいている。比較のために、我々はまた、異なる注意メカニズムを持つ2つの他のベースラインを構築する:(1)変形可能な注意を置き換えるためにグローバル注意を使用する;(2)各クエリを周囲の局所領域ではなく、その参照点のみと相互作用させる、これは以前の方法[36, 37]と同様である。より広範な比較のために、BEVFormer を VPN [30] と Lift-Splat [32] によって提案された BEV 生成手法に置き換えた。

表5: nuScenes valセットにおける、様々なBEVエンコーダを用いた様々な手法の検出結果。「メモリ」は学習時に消費されるGPUメモリである。*: 比較のために、VPN[30]とLiftSplat[32]を使用して、我々のモデルのBEVエンコーダを置き換える。†: BEVFormer-Sは空間交差注意の大域的注意を用いて学習し、モデルはfp16の重みで学習する。また、バックボーンからシングルスケールの特徴のみを採用し、BEVクエリの空間形状を 100×100 に設定することで、メモリを節約している。‡: 予測されるオフセットと重みを除去することで、局所領域から参照点のみへの変形可能な注意の相互作用ターゲットを劣化させる。

Method	Attention	NDS↑	mAP↑	mATE↓	mAOE↓	#Param.	FLOPs	Memory
VPN* [30]	-	0.334	0.252	0.926	0.598	111.2M	924.5G	~20G
List-Splat* [32]	-	0.397	0.348	0.784	0.537	74.0M	1087.7G	~20G
BEVFormer-S [†]	Global	0.404	0.325	0.837	0.442	62.1M	1245.1G	~36G
BEVFormer-S [‡]	Points	0.423	0.351	0.753	0.442	68.1M	1264.3G	~20G
BEVFormer-S	Local	0.448	0.375	0.725	0.391	68.7M	1303.5G	~20G

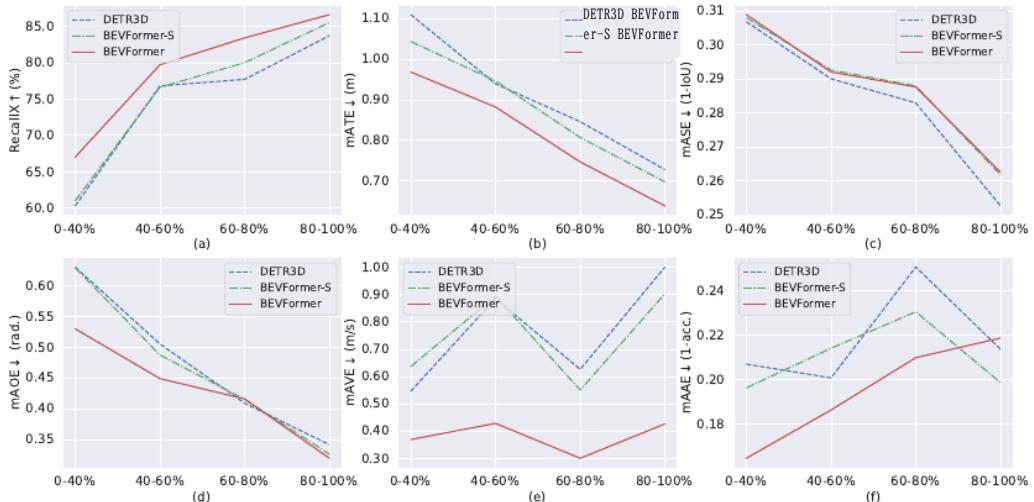


図3: 視認性の異なるサブセットの検出結果。nuScenes val setを、{0-40%, 40-60%, 60-80%, 80-100%}のオブジェクトが見えるという可視性に基づいて、4つのサブセットに分割する。(a): 時間情報によって強化されたBEVFormerは、すべてのサブセットで、特に可視性が最も低いサブセット(0-40%)で高いリコールを持つ。(b)、(d)、(e): 時間情報は並進、方位、速度の精度に有利である。(c)と(f): (c)と(f):異なる手法間のスケールと属性の誤差は最小である。時間情報はオブジェクトのスケール予測に役立つとは限らない。

変形可能な注意は、同等のモデルスケールの下で、他の注意メカニズムを大幅に上回ることが観察できる。グローバルアテンションはGPUメモリを消費しすぎ、ポイントインタラクションは受容野が限られている。スペースアテンションは、事前に決定された関心領域と相互作用し、受容野とGPU消費のバランスをとるため、より良いパフォーマンスを達成する。

時間的自己アテンションの有効性Tab. 1とTab. 4から、BEVFormerは同じ設定、特に困難な検出タスクにおいて、顕著な改善でBEVFormer-Sを上回ることが観察できる。時間情報の効果は主に以下の側面にある:(1)時間情報の導入は速度推定の精度に大きく寄与する、(2)時間情報があれば、予測される物体の位置や向きはより正確である、(3)図3に示すように、時間情報には過去の物体の手がかりが含まれているため、大きくオクルージョンした物体に対して高い再現率を得ることができる。オクルージョンレベルの異なる物体に対するBEVFormerの性能を評価するために、以下のように分割する。nuScenesの検証セットを、nuScenesが提供する公式の可視性ラベルに従って、4つのサブセットに分割する。各サブセットにおいて、マッチング中の中心距離閾値を2mとして、全カテゴリの平均想起率も計算する。

表6: nuScenes val setにおける異なるモデル構成のレイテンシと性能。V100 GPUでレイテンシを測定し、バックボーンはR101-DCNである。入力画像の形状は900×1600である。「MS」はマルチスケールビューの特徴を示す。

Method	Scale of BEVFormer			Latency (ms)			FPS	NDS↑	mAP↑
	MS	BEV	#Layer	Backbone	BEVFormer	Head			
BEVFormer	✓	200×200	6	391	130	19	1.7	0.517	0.416
A	✗	200×200	6	387	87	19	1.9	0.511	0.406
B	✓	100×100	6	391	53	18	2.0	0.504	0.402
C	✓	200×200	1	391	25	19	2.1	0.501	0.396
D	✗	100×100	1	387	7	18	2.3	0.478	0.374

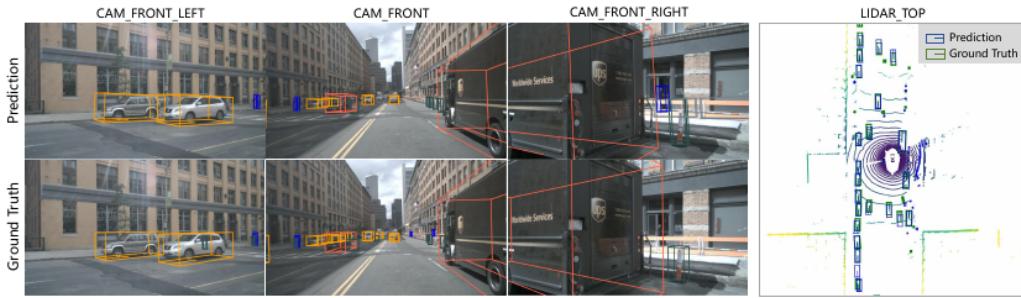


図4: nuScenes valセットに対するBEVFormerの可視化結果。マルチカメラ画像と鳥瞰図における3D bbxes予測値を示す。

予測されたボックスの最大数は、想起を公平に比較するために、すべての方法で300である。オブジェクトが0~40%しか見えないサブセットでは、BEVFormerの平均リコールはBEVFormer-SとDETR3D Truthを6.0%以上のマージンで上回る。

地上モデルのスケールとレイテンシ。表6では、異なる構成の性能とレイテンシを比較する。6. BEVFormerのスケールを、マルチスケールビュー特徴を使用するかどうか、BEVクエリの形状、レイヤー数を含む3つの側面から除去し、性能と推論レイテンシのトレードオフを検証する。BEVFormerの1つのエンコーダ層を使用する構成Cは、50.1 %のNDSを達成し、BEVFormerのレイテンシを元の130msから25msに短縮することが観察できる。シングルスケールのビュー特徴、より小さなBEVサイズ、および1つのエンコーダ層のみを持つ構成Dは、デフォルトの構成と比較して3.9ポイントを失うものの、推論中にわずか7msしか消費しない。しかし、マルチビュー画像入力のため、効率を制限するボトルネックはバックボーンにあり、自律走行のための効率的なバックボーンは詳細な研究に値する。全体として、我々のアーキテクチャは様々なモデルスケールに適応することができ、性能と効率をトレードオフする柔軟性を持つ。

4.6 可視化結果

図4に複雑なシーンの検出結果を示す。BEVFormerは、小さな物体や離れた物体でのいくつかのミスを除いて、印象的な結果を生成する。より定性的な結果は付録に記載されている。

5 考察と結論

本研究では、マルチカメラ入力から鳥瞰特徴量を生成するBEVFormerを提案した。BEVFormerは、空間情報と時間情報を効率的に集約し、3D検出と地図分割のタスクを同時にサポートする強力なBEV特徴を生成することができます。

限界。現時点では、カメラベースの手法は、LiDARベースの手法と効果と効率において、まだ特にギャップがある。2次元情報から3次元位置を正確に推定することは、カメラベースの手法にとって長年の課題である。

より広範な影響 BEVFormerは、マルチカメラ入力からの時空間情報を使用することで、視覚知覚モデルの性能を大幅に改善できることを実証している。BEVFormerが示す、より正確な速度推定や、視認性の低い物体に対する高い再現性などの利点は、より優れた、より安全な自律走行システムを構築するために不可欠である。我々は、BEVFormerが以下のようなより強力な視覚知覚手法のベースラインに過ぎず、視覚に基づく知覚システムはまだ研究すべき途方もない可能性を秘めていると考えている。

References

- [1] Brazil, G., Liu, X.: M3d-rpn: Monocular 3d region proposal network for object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9287–9296 (2019)
- [2] Brazil, G., Pons-Moll, G., Liu, X., Schiele, B.: Kinematic 3d object detection in monocular video. In: European Conference on Computer Vision. pp. 135–152. Springer (2020)
- [3] Bruls, T., Porav, H., Kunze, L., Newman, P.: The right (angled) perspective: Improving the understanding of road scenes using boosted inverse perspective mapping. In: 2019 IEEE Intelligent Vehicles Symposium (IV). pp. 302–309. IEEE (2019)
- [4] Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Lioung, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11621–11631 (2020)
- [5] Can, Y.B., Liniger, A., Paudel, D.P., Van Gool, L.: Structured bird's-eye-view traffic scene understanding from onboard images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15661–15670 (2021)
- [6] Can, Y.B., Liniger, A., Unal, O., Paudel, D., Van Gool, L.: Understanding bird's-eye view semantic hd-maps using an onboard monocular camera. arXiv preprint arXiv:2012.03040 (2020)
- [7] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
- [8] Chen, X., Ma, H., Wan, J., Li, B., Xia, T.: Multi-view 3d object detection network for autonomous driving. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 1907–1915 (2017)
- [9] Chitta, K., Prakash, A., Geiger, A.: Neat: Neural attention fields for end-to-end autonomous driving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15793–15803 (2021)
- [10] Cho, K., Van Merriënboer, B., Bahdanau, D., Bengio, Y.: ニューラル機械翻訳の特性について：エンコーダ・デコーダのアプローチarXivプレプリント arXiv:1409.1259 (2014)
- [11] Crawshaw, M.: ディープニューラルネットワークによるマルチタスク学習：サーベイ。arXivプレプリント arXiv:2009.09796 (2020)
- [12] Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.: Deformable convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp. 764–773 (2017)
- [13] 50, C., Amid, E., Zhao, Z., Yu, T., Anil, R., Finn, C.: マルチタスク学習のためのタスクグループ分けの効率的な特定。神経情報処理システムの進歩 34 (2021)
- [14] Gehring, J., Auli, M., Grangier, D., Yarats, D., Dauphin, Y.N.: Convolutional sequence to sequence learning. In: International Conference on Machine Learning. pp. 1243–1252. PMLR (2017)
- [15] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
- [16] Hendy, N., Sloan, C., Tian, F., Duan, P., Charchut, N., Xie, Y., Wang, C., Philbin, J.: Fishing net: Future inference of semantic heatmaps in grids. arXiv preprint arXiv:2006.09917 (2020)
- [17] Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation 9(8), 1735–1780 (1997)

- [18] Hu, A., Murez, Z., Mohan, N., Dudas, S., Hawke, J., Badrinarayanan, V., Cipolla, R., Kendall, A.: Fiery: In: IEEE/CVF国際コンピュータビジョン会議論文集. 15273–15282頁
- [19] Kang, K., Ouyang, W., Li, H., Wang, X.: Object detection from video tubelets with convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 817–825 (2016)
- [20] Lang, A.H., Vora, S., Caesar, H., Zhou, L., Yang, J., Beijbom, O.: Pointpillars: Fast encoders for object detection from point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12697–12705 (2019)
- [21] Lee, Y., Hwang, J.w., Lee, S., Bae, Y., Park, J.: An energy and gpu-computation efficient backbone network for real-time object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019)
- [22] Li, Z., Wang, W., Xie, E., Yu, Z., Anandkumar, A., Alvarez, J.M., Lu, T., Luo, P.: Panoptic segformer: Delving deeper into panoptic segmentation with transformers. arXiv preprint arXiv:2109.03814 (2021)
- [23] Lin, T.Y., Dollár, P., Girshick, R.B., He, K., Hariharan, B., Belongie, S.J.: Feature pyramid networks for object detection. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 936–944 (2017)
- [24] Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv: Learning (2017)
- [25] Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019)
- [26] Luo, W., Yang, B., Urtasun, R.: Fast and furious: Real time end-to-end 3d detection, tracking and motion forecasting with a single convolutional net. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 3569–3577 (2018)
- [27] Ma, X., Ouyang, W., Simonelli, A., Ricci, E.: 3d object detection from images for autonomous driving: A survey. arXiv preprint arXiv:2202.02980 (2022)
- [28] Mousavian, A., Anguelov, D., Flynn, J., Kosecka, J.: 3d bounding box estimation using deep learning and geometry. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 7074–7082 (2017)
- [29] Ng, M.H., Radia, K., Chen, J., Wang, D., Gog, I., Gonzalez, J.E.: Bev-seg: Bird’s eye view semantic segmentation using geometry and semantic point cloud. arXiv preprint arXiv:2006.11436 (2020)
- [30] Pan, B., Sun, J., Leung, H.Y.T., Andonian, A., Zhou, B.: Cross-view semantic segmentation for sensing surroundings. IEEE Robotics and Automation Letters **5**(3), 4867–4873 (2020)
- [31] Park, D., Ambrus, R., Guizilini, V., Li, J., Gaidon, A.: Is pseudo-lidar needed for monocular 3d object detection? In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3142–3152 (2021)
- [32] Philion, J., Fidler, S.: Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In: European Conference on Computer Vision. pp. 194–210. Springer (2020)
- [33] Qi, C.R., Zhou, Y., Najibi, M., Sun, P., Vo, K., Deng, B., Anguelov, D.: Offboard 3d object detection from point cloud sequences. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6134–6144 (2021)
- [34] Reading, C., Harakeh, A., Chae, J., Waslander, S.L.: Categorical depth distribution network for monocular 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8555–8564 (2021)
- [35] Reiher, L., Lampe, B., Eckstein, L.: A sim2real deep learning approach for the transformation of images from multiple vehicle-mounted cameras to a semantically segmented image in bird’s eye view. In: 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC). pp. 1–7. IEEE (2020)
- [36] Roddick, T., Kendall, A., Cipolla, R.: Orthographic feature transform for monocular 3d object detection. In: BMVC (2019)

- [37] Rukhovich, D., Vorontsova, A., Konushin, A.: Imvoxelnet: 単眼および多視点汎用3次元物体検出のための画像からボクセルへの投影。In: コンピュータビジョンの応用に関するIEEE/CVF冬季大会予稿集. 2 397–2406頁
- [38] Saha, A., Maldonado, O.M., Russell, C., Bowden, R.: Translating images into maps. arXiv preprint arXiv:2110.00966 (2021)
- [39] Simonelli, A., Bulo, S.R., Porzi, L., Lopez-Antequera, M., Kotschieder, P.: Disentangling monocular 3d object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)
- [40] Sun, P.、Kretzschmar, H.、Dotiwalla, X.、Chouard, A.、Patnaik, V.、Tsui, P.、Guo, J.、Zhou, Y.、Chi, Y.、Caine, B.、他: 自律走行のための知覚におけるスケーラビリティ: Waymoオープンデータセット。In: コンピュータビジョンとパターン認識に関するIEEE/CVF会議論文集. 2446–2454頁(2020年)
- [41] Tian, Z., Shen, C., Chen, H., He, T.: Fcos: Fully convolutional one-stage object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9627–9636 (2019)
- [42] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
- [43] Vora, S., Lang, A.H., Helou, B., Beijbom, O.: Pointpainting: Sequential fusion for 3d object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4604–4612 (2020)
- [44] Wang, T., Xinge, Z., Pang, J., Lin, D.: Probabilistic and geometric depth: Detecting objects in perspective. In: Conference on Robot Learning. pp. 1475–1485. PMLR (2022)
- [45] Wang, T., Zhu, X., Pang, J., Lin, D.: Fcos3d: Fully convolutional one-stage monocular 3d object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 913–922 (2021)
- [46] Wang, Y., Chao, W.L., Garg, D., Hariharan, B., Campbell, M., Weinberger, K.Q.: Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8445–8453 (2019)
- [47] Wang, Y., Guizilini, V.C., Zhang, T., Wang, Y., Zhao, H., Solomon, J.: Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In: Conference on Robot Learning. pp. 180–191. PMLR (2022)
- [48] Xie, E., Yu, Z., Zhou, D., Phlion, J., Anandkumar, A., Fidler, S., Luo, P., Alvarez, J.M.: M²bev: Multi-camera joint 3d detection and segmentation with unified birds-eye view representation. arXiv preprint arXiv:2204.05088 (2022)
- [49] Xu, B., Chen, Z.: Multi-level fusion based 3d object detection from monocular images. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2345–2353 (2018)
- [50] Yan, Y., Mao, Y., Li, B.: Second: Sparsely embedded convolutional detection. Sensors **18**(10), 3337 (2018)
- [51] Yang, W., Li, Q., Liu, W., Yu, Y., Ma, Y., He, S., Pan, J.: Projecting your view attentively: Monocular road scene layout estimation via cross-view transformation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15536–15545 (2021)
- [52] Yin, T., Zhou, X., Krähenbühl, P.: Center-based 3d object detection and tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11784–11793 (2021)
- [53] Zhou, X., Wang, D., Krähenbühl, P.: Objects as points. arXiv preprint arXiv:1904.07850 (2019)
- [54] Zhou, Y., Tuzel, O.: Voxelnet: End-to-end learning for point cloud based 3d object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4490–4499 (2018)
- [55] Zhu, X., Ma, Y., Wang, T., Xu, Y., Shi, J., Lin, D.: Ssn: Shape signature networks for multi-class object detection from point clouds. In: European Conference on Computer Vision. pp. 581–597. Springer (2020)

- [56] Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. In: International Conference on Learning Representations (2020)
- [57] Zhu, X., Xiong, Y., Dai, J., Yuan, L., Wei, Y.: Deep feature flow for video recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2349–2358 (2017)

Appendix

A 実装の詳細

本節では、提案手法の実装の詳細と実験について述べる。

A.1 Traning Strategy

先行手法[47, 56]に従い、全てのモデルを24エポック、GPUあたりのバッチサイズ1(6ビュー画像を含む)、学習率 2×10^{-4} 、バックボーンの学習率乗数は0.1、コサインアニーリング[24]で学習率を減衰させる。AdamW[25]を用い、重み減衰を 1×10^{-2} としてモデルを最適化する。

A.2 VPNとリフトスプラット

本研究では、VPN[30]とLift-Splat[32]を2つのベースラインとして使用する。バックボーンとタスクヘッドは、公平な比較のためにBEVFormerと同じである。

VPN。本研究では公式コード¹を採用する。MLPの膨大なパラメータに制限され、VPNが高解像度のBEV(例えば 200×200)を生成することは困難である。VPNと比較するために、本研究では、2つのビュー変換層を介して、シングルスケールのビュー特徴を 50×50 の低解像度でBEVに変換する。

Lift-Splat。Lift-Splat²のカメラエンコーダを2つの畳み込み層を追加して拡張し、同等のパラメータ数で我々のBEVFormerと公平に比較する。他の設定は変更しない。

A.3 Task Heads

検出ヘッド。各3次元バウンディングボックスに対して、各ボックスのスケールを表す3つのパラメータ(l, w, h)、中心位置を表す3つのパラメータ(x_o, y_o, z_o)、物体のヨー $\text{Reshape } \theta$ を表す2つのパラメータ($\cos(\theta), \sin(\theta)$)、速度を表す2つのパラメータ(v_x, v_y)を含む10個のパラメータを予測する。 L_1 損失と L_1 コスト_{BEV}のみが_{Features}学習View段階で使用される。[47]に従い、900のオブジェクトクエリを使用し、 256×28 の推論中、最も信頼スコアの高い300の予測ボックスを保持する。 $\times 50_{256 \times 50 \times 50}$

セマンティーションヘッド。 256×1400 図5に示すように、セマンティックマップの各クラスについて、[22]のマスクデコーダに従い、このクラスを表現するために学習可能なクエリを1つ使用し、バニラマルチヘッドアテンションからの 256 アテンション $\times 2500$ マップに基づいて $\times 2500$ 最終的なセマンティーションマスク 256 を生成する。

A.4 空間的交差注意

グローバルな注意。変形可能な注意[56]の他に、我々の空間的交差注意はグローバル注意(すなわち、バニラ多頭注意)[42]によっても実装できる。グローバルアテンションを採用する最も簡単な方法は、各BEVクエリをすべてのマルチカメラ特徴と相互作用させることであり、この概念的な実装はカメラキャリブレーションを必要としない。しかし、この素直な方法の計算コストは手が出ない。したがって、カメラのintrinsicとextrinsicを利用して、1つのBEVクエリが対話に値するヒットビューを決定する。この戦略により、1つのBEVクエリは通常、すべてのビューではなく、1つか2つのビューのみと相互作用し、空間的交差注意にグローバルな注意を使用することが可能になる。注目すべきは、正確なカメラの内在的・外在的に依存している他の注意メカニズムと比較して、グローバル注意はカメラのキャリブレーションに対してより頑健であることである。

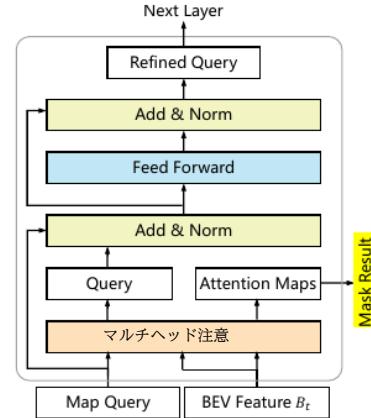


図5:セグメンテーション(b) BEVFormerのマスクデコーダヘッド(マスクデコーダ)。

¹ <https://github.com/pbw-Berwin/View-Parsing-Network>
² <https://github.com/nv-tlabs/lift-splat-shoot>

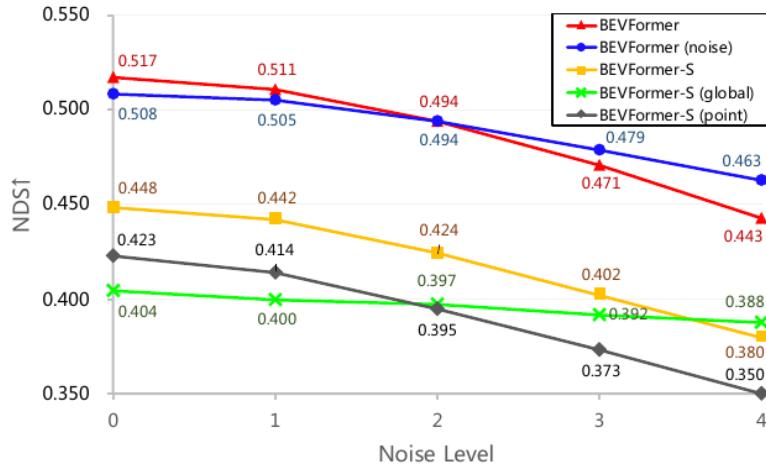


図6:異なるレベルのカメラ外部ノイズにさらされたnuScenes valセットに対する手法のNDS。i番目のレベルのノイズについては、回転ノイズは平均が0、分散が*i*に等しい正規分布からサンプリングされ(回転ノイズは度数で、各軸のノイズは独立)、並進ノイズは平均が0、分散が $5i$ に等しい正規分布からサンプリングされる(並進ノイズはセンチメートルで、各方向のノイズは独立)。「BEVFormer」はデフォルトのバージョンである。「BEVFormer(ノイズ)」はノイズの多いエクストリンシック(ノイズレベル=1)で学習される。「BEVFormer-S」は、変形可能な注意[56]によって実装された空間交差注意を持つ、我々の静的バージョンのBEVFormerである。「BEVFormer-S(global)」は、グローバルアテンション(すなわち、バニラマルチヘッドアテンション)によって実装された空間クロスアテンションを持つBEVFormer-Sである[42]。「BEVFormer-S(点)」はBEVFormer-S with point spatial cross-attentionであり、予測されるオフセットと重みを除去することで、局所領域から参照点のみへの変形可能な注意の相互作用ターゲットを劣化させる。

B カメラ外挿に対するロバスト性

BEVFormerは、2Dビューの基準点を得るために、カメラのイントリンシックとエクストリンシックに依存する。自律走行システムの展開段階では、キャリブレーションエラーやカメラオフセットなど、様々な理由でエクストリンシックに偏りが生じる可能性がある。図6に示すように、カメラのエクストリンシックのノイズレベルが異なる場合のモデルの結果を示す。BEVFormer-S(点)と比較して、BEVFormer-Sは変形可能な注意[56]に基づく空間交差注意を利用し、参照点のみと相互作用するのではなく、参照点周辺の特徴をサンプリングする。変形可能な注意により、BEVFormer-Sのロバスト性はBEVFormer-S(点)よりも強くなる。例えば、ノイズレベルが4の場合、BEVFormer-SのNDSは15.2%($1 - \frac{0.380}{0.448}$ で計算)低下し、BEVFormer-S(点)のNDSは17.3%低下する。BEVFormer-Sと比較すると、BEVFormerは14.3%のNDSしか低下させておらず、これは時間情報がカメラ外挿に対するロバスト性も向上させることを示している。[32]に従い、ノイズの多いエクストリンシックでBEVFormerを学習させた場合、BEVFormer(ノイズ)の方がロバスト性が高いことを示す(NDSは8.9%しか低下しない)。グローバルアテンションに基づく空間交差アテンションにより、BEVFormer(global)はカメラ外挿ノイズのレベル4でも強い干渉防止能力(NDS低下4.0%)を持つ。その理由は、BEVクリエイのROIを選択するためにカメラエクストリンシクスを利用しないためである。

注目すべきは、最も厳しいノイズの下で、BEVFormer-S(global)がBEVFormer-Sを上回ることさえあることがある(38.8% NDS対38.0% NDS)。

C アブレーション研究

学習時のフレーム数の影響。表7に学習時のフレーム数の影響を示す。nuScenes valセットのNDSはフレーム数の増加とともに上昇を続け、フレーム数 ≥ 4 から横ばいになり始めていることがわかる。そこで、実験ではデフォルトで学習時のフレーム数を4に設定した。

表7: nuScenes valセットにおける、学習時のフレーム数を変えたモデルのNDS。”#Frame”は学習時のフレーム番号を表す。

#Frame	NDS↑	mAP↑	mAve↓
1	0.448	0.375	0.802
2	0.490	0.388	0.467
3	0.510	0.410	0.423
4	0.517	0.416	0.394
5	0.517	0.412	0.387

表8: nuScenes valセットでのアプレーション実験。”A.”は、履歴BEVの特徴を自我運動と整合させることを示す。”R.”は、5つの連続フレームから4フレームをランダムにサンプリングすることを示す。”B.”は、オフセットと重みを予測するために、BEVクエリと履歴BEV特徴の両方を使用することを示す。

#	A.	R.	B.	NDS↑	mAP↑
1	✗	✓	✓	0.510	0.410
2	✓	✗	✓	0.513	0.410
3	✓	✓	✗	0.513	0.404
4	✓	✓	✓	0.517	0.416

いくつかのデザインの効果。Table. 8は、いくつかのアプレーション研究の結果を示している。1と4を比較すると、現在のBEVクエリと同じジオメトリシーンを表現するためには、履歴BEV特徴をエゴモーションに合わせることが重要であることがわかる(NDS 51.0%対NDS 51.7%)。#2と#4を比較すると、5フレームから4フレームをランダムにサンプリングすることは、性能を向上させる効果的なデータ増強戦略である(NDS 51.3%対NDS 51.7%)。時間的自己注意モジュール(3参照)において、オフセットと重みを予測するためにBEVクエリのみを使用する場合と比較して、BEVクエリと履歴BEV特徴(4参照)の両方を使用する場合は、過去のBEV特徴に関する手がかりが多く、位置予測に有利です(NDS 51.3% vs. NDS 51.7%)。

D Visualization

図7に示すように、BEVFormerとBEVFormer-Sを比較する。時間情報により、BEVFormerはボーダーに遮られた2台のバスを検出することに成功した。また、物体検出と地図分割の結果を図8に示すが、検出結果と分割結果は非常に一致していることがわかる。図9に、より多くのマップ分割結果を示す。BEVFormerによって生成された強力なBEV特徴により、単純なマスクデコーダを介して意味マップをうまく予測できることがわかる。

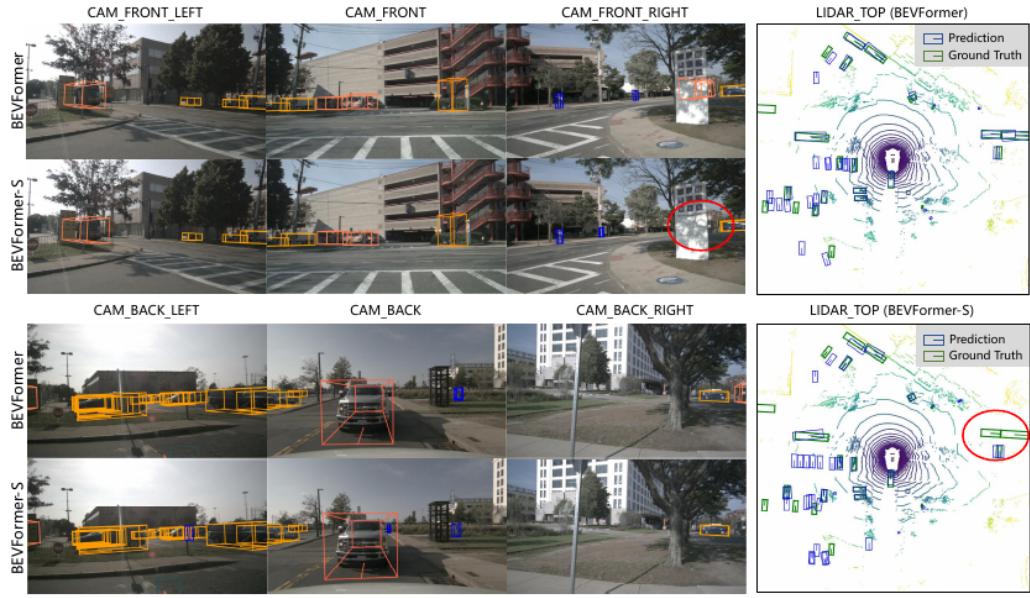


図7: nuScenes valセットにおけるBEVFormerとBEVFormer-Sの比較。BEVFormerは高度に隠蔽された物体を検出することができ、これらの物体はBEVFormer-Sの予測結果(赤丸)では見逃されていることがわかる。

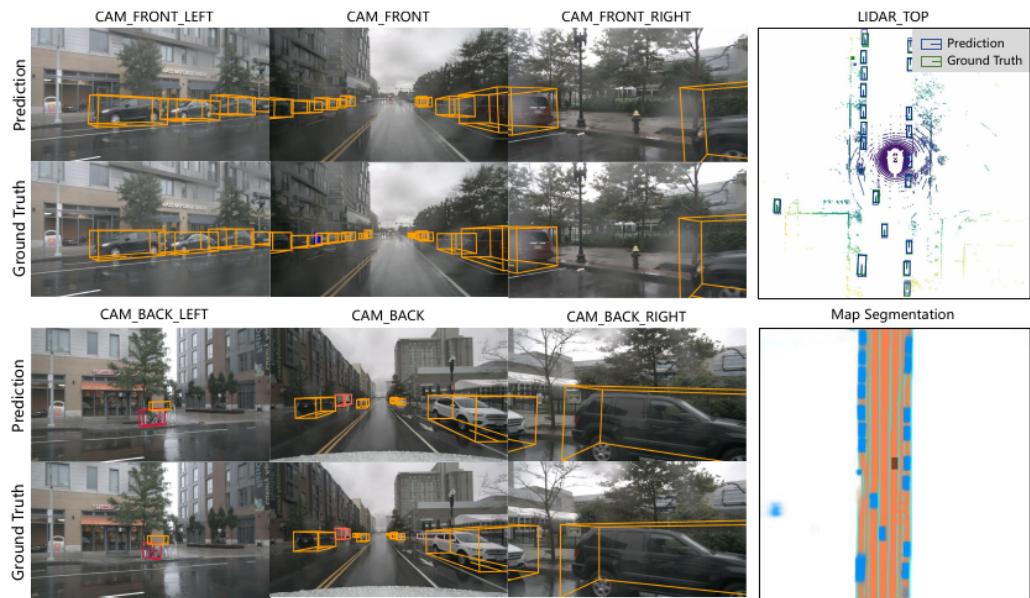


図8: 物体検出とマップセグメンテーションの両タスクの可視化結果。車両、道路、車線のセグメンテーションをそれぞれ青、オレンジ、緑で示す。

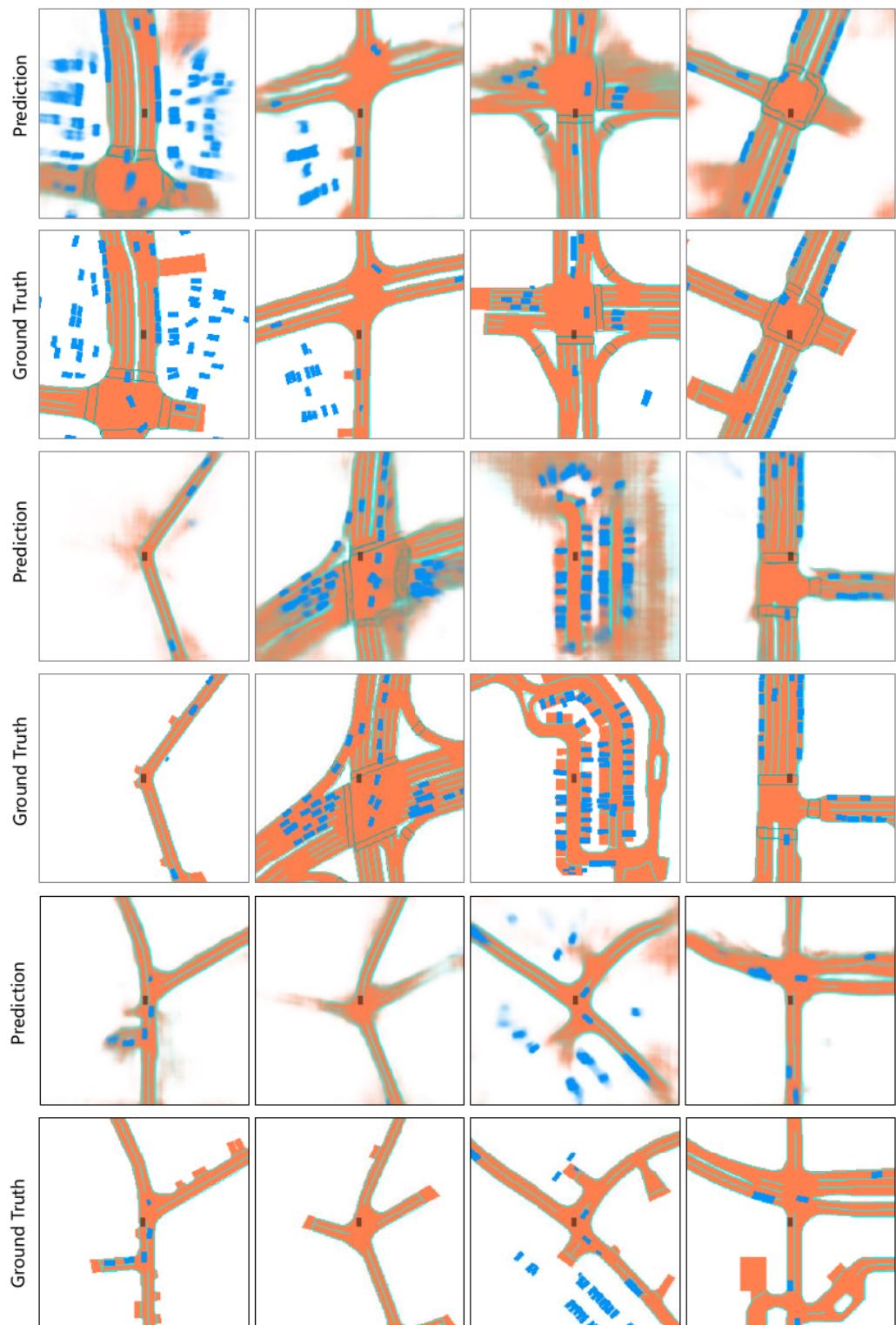


図9:マップセグメンテーションタスクの可視化結果。車両、道路、歩行横断、車線分割をそれぞれ青、オレンジ、シアン、緑で示す。